Not Directly Stated, Not Explicitly Stored:

Conversational Agents and the Privacy Threat of Implicit Information

Martha Larson
Institute for Computing and Information Sciences and Centre for Language Studies, Radboud University, Netherlands
m.larson@cs.ru.nl

Nelleke Oostdijk
Centre for Language Studies, Radboud University, Netherlands
n.oostdijk@let.ru.nl

Frederik Zuiderveen Borgesius
Interdisciplinary Hub for Security, Privacy and Data Governance, Radboud University, Netherlands
frederikzb@cs.ru.nl

ABSTRACT

As conversational agents continue to evolve, it will become increasingly common to interact with search engines and recommender systems via natural language dialogue. Such interactions guide and shape our decision making, especially our consumption of products and services. The evolution of conversational agents will bring new challenges in protecting the privacy of users and research has already begun to identify and address potential threats. Current research, however, focuses on how conversational agents acquire and process explicit information. In this paper, we consider the future and bring to light the up-and-coming privacy risks posed by implicit information. Our first point is that meaning that is expressed implicitly is an integral part of natural language, implying that agents that have the ability to engage in a fully humanlike dialogue will also have the ability to manipulate implied meaning. As a result, such agents will be capable of acquiring sensitive information about users that is not directly stated. Users have little awareness of or control over information that is implicitly communicated. Our second point is that in today's search and recommender systems user profiles are not explicitly stored. As a result, it is not obvious that a user is being targeted on the basis of implicit person-specific information. The way forward, we argue, is for research in the area of conversational agents to devote more attention to the linguistic principles that underlie implied meaning and the legal means that are available to protect users.

CCS CONCEPTS

• **Security and privacy** → Human and societal aspects of security and privacy; Social aspects of security and privacy; • **Computing methodologies** → Artificial intelligence; Natural language processing; Discourse, dialogue and pragmatics.

KEYWORDS

Conversational agents, Privacy, Data protection law, Linguistics, Pragmatics

UMAP '21 Adjunct, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8367-7/21/06.
https://doi.org/10.1145/3450614.3463601

ACM Reference Format:
Martha Larson, Nelleke Oostdijk, and Frederik Zuiderveen Borgesius. 2021. Not Directly Stated, Not Explicitly Stored:: Conversational Agents and the Privacy Threat of Implicit Information. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3450614.3463601

1 INTRODUCTION

We are moving towards a future in which conversational agents are ubiquitous and can engage people in dialogues that are fully humanlike. Interaction with search engines and recommender systems will increasingly take the form of natural language exchange. In this future, conversational agents will guide and shape our decision making, especially our consumption of products and services. The prospect of this future raises important concerns about privacy. Although a growing amount of research is aimed at addressing these concerns, we find that current research on conversational agents is missing an important privacy threat: *implicit information*. Currently, existing work focuses on *explicit information*, which is directly stated in the dialogue and explicitly stored by the conversational system. Specifically, researchers have studied issues related to the collection and processing of data [13] [1] [15]. They have also investigated whether human-like (anthropomorphic) voice assistants could lead people to reveal more than they intend [11]. Recently, researchers have begun to draw attention to the issues of privacy and conversational agents that are related to explicit information [10]. In all this work, the focus is on explicit information. Moving forward, the danger is that system designers and regulators will work to protect explicit information but will overlook privacy-sensitive information that is *not directly stated* and *not explicitly stored*. The goal of this paper is to bring to light the up-and-coming privacy risks posed by implicit information in the area of conversational agents.

2 ANTICIPATING THE RISKS OF INFORMATION THAT IS NOT DIRECTLY STATED

In this section, we explain how people use natural language to communicate and control implied meaning and discuss the implications for conversational agents. Implied meaning is studied in the field of linguistics known as *pragmatics* [8] [9]. In the past, linguists have teamed up with legal experts to understand how people manipulate implied meaning in situations in which language has legal implications. For example, researchers have described how lawyers in the

courtroom control the ability of a witness under examination to use pragmatic strategies to avoid answering a question directly [2]. Further, researchers have documented how pragmatics is abused during police interrogation with the effect of circumventing the law and coercing confession [5]. Our concern is that advances in natural language capacity could turn conversational agents into highly trained cross-examiners or interrogators, whose ability to manipulate implied meaning is so subtle that it goes undetected. Linguists and legal experts must collaborate with artificial intelligence (AI) experts to explore the potential of this threat.

To grasp the risk of conversational agents that can manipulate implied meaning, consider the following dialogue between an AI-based conversational system (virtual assistant) and a father trying to buy a birthday present for his daughter without revealing private information.

Agent: How can I help you?
Father: I am looking for a fun present for a kid.
Agent: Is it for your son or daughter?
Father: Just a fun present for any kid.
Agent: What is the child's age?
Father: A general present for a kid is fine.
Agent: How about a trampoline?

The conversational agent asks direct questions, and the father does not answer them. He has successfully avoided sharing details about the gender of the child, whether or not it is his child, and the age of the child. Now consider, however, how this dialogue continues:

Father: Aren't trampolines kind of dangerous?
Agent: *What specifically are you worried about?*
Father: I'm worried the kid will hurt herself.
Agent: *Our trampolines are all independently certified.*
Father: I don't think she's old enough yet.
Agent: *Why not consider a starter trampoline, then?*
Father: My wife worries about the kids. Can you suggest something safe?

The conversational agent has stopped asking direct questions about the child, and the conversation seems to focus on the trampoline. However, due to the principles of pragmatics, the user 'leaks' the information about the child that he was previously trying to withhold. To see this, note the following: The sentence 'I'm worried the kid will hurt. . .' cannot be naturally finished without a pronoun that reveals the gender of the child. 'My wife worries about the kids.' implies that the child in question is his own, and 'I don't think she's old enough yet' suggests approximate information about the age of the child. In short, the father has made no direct statements providing information about his family, nor has the conversational agent made any direct requests for such, and yet the conversational AI is able to collect this information anyway. This example is hypothetical; however, the privacy risk is clear. Note that the conversational agent is not particularly friendly or humanlike in this example: current work on anthropomorphism falls far short of capturing the full privacy risk of voice assistants.

Today, a growing number of conversational agents are not directly programmed but rather are based on machine learning techniques. Essentially, such conversational agents are a type of AI that 'learns' how to interact by finding patterns in large amounts of data. The learning process is driven by measurable objectives defined by the agent's designers. The objective of a voice assistant might be to suggest a product that the user buys and does not afterwards return. On the surface, the objectives are benign. However, problems may arise because the workings of the underlying algorithms dictate that the agent must pursue its objectives relentlessly. Consequently, if the use of detailed person-specific information to make product suggestions leads the user to make a purchase and not return it, then the agent will persistently pursue the acquisition of this information. If the agent is blocked from asking obvious, direct questions, then it may learn paths to acquire the information it needs indirectly, subtly, and unnoticeably. The agent can accumulate information from small leaks over time to acquire substantial, privacy-sensitive information.

3 ANTICIPATING THE RISKS OF INFORMATION THAT IS NOT EXPLICITLY STORED

Now that we have explained how conversational agents could control natural language to collect implicit information, without the need for the information to ever be directly stated in the dialogue, we turn to issues related to information storage. One possible way to address the privacy threat of implicit information would be for designers and regulators to check the information that is stored by the conversational agent after the dialogue, and to delete all privacy-sensitive information that has been accumulated. Such a solution comes into consideration for systems that explicitly record the information that they have inferred from the dialogue, such as is discussed in [10]. However, not all systems store the information gathered from users in an explicit form. In fact, the state of the art is moving in the direction of systems that store user representations that can be used as the basis of personalization (i.e., making future recommendations), but that are not useful for people. In other words, a person who looked at the representation would not be able to interpret it and could not tell whether or not it contained privacy-sensitive information. A representative example is the conversational recommender system in [4]. The system asks the user explicit questions, but under the hood uses probabilistic matrix factorization, a technique that does not create human-readable representations.

Another issue is that there might not be a moment in the dialogue at which the information could be reviewed. The machine-learning-based AI used by recommender systems jumps directly from raw data to action (e.g., offering an income-specific price without explicitly leveraging income information). This jump means that deleting user representations after the conversation session is not enough to prevent users from being targeted on the basis of information that they did not intend to share. For example, in the dialogue above, the system could use the information that the child is a girl to steer the father towards gender specific toys. It is plausible that the system would do this, since gendered marketing can increase sales [14]. However, such targeting represents an interference with the prerogative of the father to determine the degree to which his child's upbringing is gender specific.

It is important to realize that designers of conversational AI choose to use machine learning because they cannot fully anticipate

or understand which variables are important. The fact that internal representations of users are not human readable is not a 'bug' but is actually a function of how AI is intended to work.

4 THE ROAD AHEAD

With this paper, we aim to raise awareness of the privacy threats of conversational agents involving implicit information. Although current systems may not yet present these risks in obvious ways, danger lies in the road ahead, as conversational agents continue to develop in sophistication. The European Data Protection Board (in which European privacy authorities cooperate) has published guidelines for virtual voice assistants [6]. These guidelines assume that user profiling for personalized content or advertising involves labels that are explicitly stored. We argue that this assumption will soon become outdated. The first evidence that language technology will be able to acquire the ability to interpret and produce implied meanings (i.e., manipulate language pragmatics) has just been published [12]. Recently, a fully autonomous debating system has been developed [16]. In this section, we provide an outlook for researchers and designers working on conversational agents to consider.

People are ill-equipped to guard themselves against conversational agents that collect implicit information. Conversational AI draws on huge amounts of training data and computational resources, making a conversational agent a superhuman interlocuter along three dimensions. First, conversational agents can instantly craft the perfect next dialogue turn, where people often suffer *l'esprit d'escalier*. Second, conversational agents can continue to engage and maneuver, whereas people will tire with time and grow less cautious. Third, conversational agents can manipulate implied meaning in a subtle way. People interacting with the agent would not necessarily be aware that their privacy, or their consumer rights, were being violated. Conversational agents exercise soft power, which coerces rather than imposes and is for this reason difficult to pin down or curb [17].

For completeness, we would like to point out that it is useful to differentiate two types of implicit information that can be derived from a conversation. The first type is associated with implied meaning, which is studied by the area of linguistics called pragmatics. The second type is associated with computational inference and would be considered more closely linked to forensic linguistics. For example, computational inference could be used to estimate someone's level of education based on their active vocabulary. Machine learning can potentially be used to acquire both types of implicit information, and a machine-learning-based conversational agent that is optimized with respect to an objective would not differentiate the two. Here, we have focused on implied meaning, because its potential for misuse by conversational AI has not yet, to our knowledge, been noticed or explored. Also, since implied meaning is an integral part of language, preventing its misuse by conversational agents will require serious research attention. Future research must investigate implicit information from the perspective of both implied meaning and computational inference.

Empirical study of the privacy threats of conversational agents is problematic. The idea of waiting until fully humanlike conversational agents have developed, and then carrying out large-scale, longitudinal studies is ill advised. For example, to understand whether a conversational agent is learning to become ageist, its interactions could be observed over a long period of time with a group of people whose ages are known. Such a study is ethically problematic, not to mention costly. Inspecting the agent's underlying model is also not readily feasible. The most recent AI that has been developed for the generation of natural human language uses 175 billion parameters [3].

Further it is important to consider the nature of the information to be protected. The categories that most easily come to mind are those that are legally considered sensitive, such as race, political affiliation, religion, sexual orientation, and state of health (e.g., Article 9 of the GDPR [7]). Categories like gender and age may be considered less sensitive, but could also be used for unfair discrimination. It is also important to protect information about people's states and traits, such as anxiety or impulsiveness; such information could be used, for instance, for manipulative targeted advertising. Users might also be interested in keeping other information confidential, for example, that they are a fashionista or an early adopter, which could cause them to be heavily targeted with a specific product category.

Research is necessary to understand how conversational AI and privacy can co-exist. Our goal has been to provide early warning of the potential privacy risks posed by implicit information in the area of conversational agents. Addressing these risks will not necessarily impede the usefulness of conversational agents. Instead, more work is necessary to develop conversational AI that can serve users without posing a threat to them. Most obviously, this work should investigate conversational AI that uses only information that is directly requested and directly stated. The work should be willing to abandon the assumption that useful personalization requires a large amount of detailed user data. For developing conversational AI that does not require large amounts of user data, inspiration can be drawn from recommender system research on using minimal necessary data.

Linguists and legal experts have a lot to offer. Pragmatics provides a framework of linguistic principles for analyzing dialogues. In this way, implied meaning can be studied by way of linguistic structures that can be directly observed in the training data of the conversational agent and in the dialogue turns it produces, rather than by way of the implied meaning that the agent has collected (which may not be explicitly represented). Legal scholars can contribute because of their familiarity with the challenge of pragmatics in situations in which language has legal implications, as mentioned above. Legal scholars can also help AI researchers to develop conversational agents that comply with legal privacy requirements (e.g., from the GDPR [7] in Europe). Consumer protection law can also be relevant, as it aims to protect people against unfair or misleading practices, such as manipulative advertising (e.g., [18] in Europe). Legal scholars can also assess whether the law leaves gaps, and whether the law should be amended to better protect people in this context. In sum, we call for more research in collaboration with linguists and legal experts to understand the potential for conversational agents to manipulate meaning and to develop regulation to prevent the misuse of implicit information.

REFERENCES

[1] Belen Sağlam, R. & Nurse, J.R.C. (2020). Is your chatbot GDPR compliant? Open issues in agent design. In Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20). ACM, Article 16, 1–3.

[2] Bianchi C. (2016). What Did You (Legally) Say? Cooperative and Strategic Interactions. In A. Capone & F. Poggi (Eds.) Pragmatics and Law. Perspectives in Pragmatics, Philosophy & Psychology, vol 7. Springer, Cham.

[3] Brown, T.B., Mann, B. Ryder, N. *et al.* (2020). Language Models are Few-Shot Learners. Neural Information Processing Systems (NeurIPS 2020).

[4] Christakopoulou, K., Filip Radlinski, F. & Hofmann, K. (2016). Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '16*). Association for Computing Machinery, New York, NY, USA, 815–824.

[5] Davis, D. & Leo, R.A. (2012). Interrogation Through Pragmatic Implication: Sticking to the Letter of the Law While Violating its Intent. In L. Solan & P. Tiersma (Eds.) Oxford Handbook on Language and the Law, Oxford University Press.

[6] European Data Protection Board. (2021). Guidelines 02/2021 on Virtual Voice Assistants, Version 1.0, 9 March 2021 (Accessed 5 April 2021)

[7] GDPR (2016) General Data Protection Regulation, Regulation (EU) 2016/679 of the European Parliament and of the Council Of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

[8] Green, Georgia M. 1989. Pragmatics and Natural Language Understanding. Lawrence Erlbaum Associates.

[9] Huang, Yan. 2014. Pragmatics. Oxford Textbooks in Linguistics. Second Edition. Oxford University Press.

[10] Hendrickx, I., van Waterschoot, J., Khan, A., ten Bosch, L., Cucchiarini, C. & Strik, H. (2021). Take Back Control: User Privacy and Transparency Concerns in Personalized Conversational Agents. Joint Proceedings of the ACM IUI 2021 Workshops.

[11] Ischen, C., Araujo, T., Voorveld, H., van Noort, G. & Smit, E. (2019). Privacy Concerns in Chatbot Interactions. In A. Følstad *et al.* (Eds.): CONVERSATIONS 2019, LNCS 11970, pp. 34–48.

[12] Jeretič, P., Warstadt, A., Bhooshan, S. & Williams, A. (2020). Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8690–8705.

[13] Pal, D., Arpnikanondt, C., Razzaque, M.A. & Funilkul, S. (2020). To Trust or Not-Trust: Privacy Issues with Voice Assistants. In IT Professional, vol. 22, no. 5, pp. 46-53, 1 Sept.-Oct. 2020.

[14] Powers, K. (2019). Shattering Gendered Marketing. American Marketing Association. 2 September 2019. https://www.ama.org/marketing-news/shattering-gendered-marketing/ (Accessed 5 April 2021)

[15] Schaub, L-P., Bruzaud, C. & Paroubek, P. (2020.) GDPR Compliance for task-oriented dialog systems conception. Proceedings of the workshop on Legal and Ethical Issues (Legal2020), Language Resources and Evaluation Conference (LREC 2020), pages 4–8.

[16] Slonim, N., *et al.* An autonomous debating system. (2021). Nature. Vol. 591, 18 March 2021 pp. 379-385.

[17] Véliz, Carissa. (2020). Privacy is Power: Why and how you should take back control of your data. Bantam Press.

[18] Unfair Commercial Practices Directive (2005) Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive').